An Empirical Study of Remote Homology Detection using Protein Language Models

Abstract—**Detecting remote homologs, proteins that share evolutionary ancestry despite low sequence similarity, remains a central challenge in computational biology. The Structural Classification of Proteins extended (SCOPe) database organizes protein domains into superfamilies based on structural and functional evidence of common origin, making it a widely used benchmark for remote homology detection. In this study, we investigate the effectiveness of protein language model (PLM) embeddings for predicting SCOPe superfamilies directly from sequence. We introduce DOMCLASS, a deep learning framework that combines supervised contrastive learning with a distance-weighted K-NN classifier to learn and exploit an embedding space aligned with SCOPe superfamily annotations. Our empirical results show that general-purpose PLM embeddings already outperform sequence similarity- and profile-based methods for remote homology detection, and that contrastive learning can further improve performance, even in challenging low sequence identity settings.**

Index Terms—**protein sequence, remote homology, protein language models, contrastive learning**

I. INTRODUCTION

Many cellular functions are carried out by proteins. These molecules are linear chains of varying lengths, built from a set of 20 standard amino acids. The specific sequence determines the protein's unique three-dimensional structure and, in turn, its function [2]. Large proteins often consist of multiple domains, which are structural units that fold relatively independently from the rest of the protein and can be understood as a modular unit of function [6]. The same domain, folded into the same three-dimensional structure, can be found in different proteins.

The principle that sequence determines structure, and structure determines function, has long guided efforts to classify protein domains into functional families based on their sequences [25], [28]. With the increasing availability of high-resolution three-dimensional (3D) structural data, this classification has extended beyond families to superfamilies. Structural information often uncovers evolutionary or functional relationships that cannot be discerned at the sequence level alone [19]. One such classification of protein domains is the Structural Classification of Proteins extended (SCOPe) [25], which organizes domains from experimentally determined structures in the Protein Data Bank (PDB). SCOPe uses expert curation to group domains that show strong structural and functional evidence of common evolutionary origin (homology) into *superfamilies*, despite having little or no detectable sequence similarity. This classification has become a gold standard for evaluating methods aimed at detecting such cases, commonly referred to as remote homology [10], [24], [33].

Traditional methods for remote homology detection rely on sequence similarity approaches, such as multiple sequence alignments, to identify evolutionary relationships between proteins. However, the sensitivity of these methods often diminishes when detecting homologous proteins with low sequence identity, often referred to as the *twilight zone* of sequence similarity [30]. Although significant advances have been made in structure-based comparisons for this problem, their application at scale is still limited by computational demands [26], [33]. Recently, transfer learning – using numerical vector representations (embeddings) derived from pre-trained protein language models (PLM) – has emerged as an alternative that could go beyond sequence-based comparisons [13], [19].

Among these, the Evolutionary Scale Modeling (ESM) family of PLMs are transformer models trained on millions of protein sequences, designed to capture a diversity of higher-level features of proteins [23], [29]. These models have shown high proficiency in various downstream tasks, such as enzyme commission number prediction [35], protein-protein contact prediction [34], and DNA-binding protein prediction [27].

In this work, we explore the application of ESM embeddings to remote homology detection by predicting SCOPe superfamilies for domain sequences. We propose a method called Domain Contrastive Learning Annotation of SuperfamilieS (DOMCLASS), a deep learning approach that refines an embedding space optimized to reflect evolutionary ancestry based on SCOPe superfamilies. A distance-weighted K-NN classifier is then used to assign superfamilies to new domains. Our results provide empirical evidence that general PLM embeddings are already more effective at identifying remote homology traditional sequence- and profile-based methods [4], [15]. We further show that contrastive learning enhances the predictive performance of PLM embeddings, particularly in challenging cases where sequence identity with annotated domains is very low – the *twilight zone* of sequence similarity.

II. RELATED WORKS

Remote homology detection remains a foundational problem in computational biology, concerned with identifying proteins that share common ancestry despite low sequence identity [10]. Traditional methods like BLASTp rely on pairwise sequence alignment to infer homology, offering fast and interpretable results but limited sensitivity in remote cases [4]. To improve in this scenario, profile-based methods based on Hidden Markov Models (HMMs) have been widely adopted [22]. These models transform multiple sequence alignments into position-specific scoring systems, enabling probabilistic comparisons between sequences and profiles [15], [20], [31]. More recently, structure-based tools such as Foldseek encoded 3D structures into discrete representations for efficient structural comparison, leveraging the fact that structure is often more conserved than sequence to search for homologs [33]. Although significant advances have been made in structure-based comparisons, their application at scale is still limited by computational demands [26], [33].

The emergence of PLMs, such as the ESM models, has enabled the creation of efficient numerical representations (embeddings) of protein sequences that capture diverse aspects of protein structure and function [29], [23]. These embeddings have shown significant success on tasks like protein subcellular localization and prediction of disease variant effects. [16], [7].

Building upon these initial representations, contrastive learning provides a powerful framework to further refine protein embeddings, optimizing an embedding space where distances correspond to biologically meaningful relationships [5]. This approach explicitly trains a model to minimize the distance between embeddings of protein sequences that belong to the same category (label) while maximizing the distance between embeddings of different ones [21]. This makes contrastive learning particularly suitable for biological datasets, which are often characterized by limited and imbalanced data, where traditional classification methods struggle due to a lack of sufficient positive examples for many categories [12]. By learning from both positive and negative examples, contrastive learning can better handle underrepresented proteins, leading to improved performance.

Contrastive learning approaches have been successfully applied to different tasks. One example is Contrastive learning-enabled enzyme annotation (CLEAN), which uses a contrastive learning framework with input from PLMs like ESM-1b to predict enzyme commission (EC) numbers [35]. CLEAN's learned embedding space reflects functional similarities, allowing it to achieve better accuracy, reliability, and sensitivity in identifying promiscuous enzymes, outperforming state-of-the-art tools based on traditional classification or sequence similarity.

Similarly, ProtTucker applies contrastive learning to protein embeddings to optimize a representation space aligned with the hierarchical classification of protein 3D structures in CATH (Class, Architecture, Topology, Homologous superfamily). ProtTucker [19] trains a feed-forward neural network on PLM embeddings using a contrastive loss based on protein triplets sampled according to the CATH hierarchy. This process leads to a learned embedding space where structural relationships are better captured than by raw embeddings or traditional sequence methods, including more distant homologies in the zone of low sequence similarity. Following these successful applications of contrastive learning to structural (CATH) and functional (EC number) hierarchies, our work aims to explore the use of PLM embeddings, specifically ESM-1b and ESM-2, combined with contrastive learning, for remote homology prediction using the SCOPe database [17].

III. Methods

A. Dataset and Data representation

SCOPe is a widely used resource that provides a manually curated hierarchical classification of protein domains based on structural and evolutionary relationships [17]. The hierarchy comprises several levels: *Family* groups domains with high sequence similarity; *Superfamily* brings together families inferred to share a common evolutionary ancestor, even in the absence of detectable sequence similarity; *Fold* clusters superfamilies with similar overall structural architecture but without strong evidence of shared ancestry; and *Class* organizes folds according to broad secondary structure content. In this study, we focus on the superfamily level as a proxy for remote homology detection that considers domains within the same superfamily to be remote homologs [10], [24], [33].

We used SCOPe versions 2.07[1] and 2.08[2] as the main resources for this study. Protein sequences were represented using embeddings derived from two pre-trained protein language models (Fig. 1a): ESM-1b [29] and ESM2-3B [23]. Each model produces a sequence of residue-level embeddings, yielding an $L \times N$ matrix for a domain of length L, where $N = 1280$ for ESM-1b and $N = 2560$ for ESM2-3B. To obtain a fixed-size representation for each domain, we compute the mean over the length dimension, resulting in a single N-dimensional embedding vector. Both models were used as static encoders, with no gradient updates or fine-tuning performed during training.

B. Learning a refined embedding space

DomCLASS uses a feedforward neural network to transform the high-dimensional embeddings produced by the PLMs into lower-dimensional embeddings. It employs a contrastive learning framework [19], [21] to learn an embedding space where Euclidean distances reflect evolutionary similarity. The network produces vectors $z \in \mathbb{R}^n$ for each domain, where spatial proximity reflects shared evolutionary origin – i.e., domains from the same superfamily are embedded closer together. The networks are trained using a Supervised Contrastive Loss [21], defined as follows:

$$\mathcal{L}^{sup} = -\sum_{d \in sf} \frac{1}{|P(d)|} \sum_{z_p \in P(d)} \log \frac{\exp(z_d \cdot z_p / \tau)}{\sum_{z_a \in A(d)} \exp(z_d \cdot z_a / \tau)} \tag{1}$$

Here, each anchor is represented z_d, corresponding to a domain with superfamily label $d \in sf$. The set $P(d)$ consists of one or more positive samples belonging to the same superfamily as the anchor. In cases where a superfamily contains only a single domain, we follow the approach of Yu et al. [35] and generate positive samples by introducing mutations into the original sequence. The set of negatives $N(d)$ includes embeddings from domains assigned to different superfamilies, but located near the anchor in the learned embedding space,

[1] https://scop.berkeley.edu/downloads/?ver=2.07

[2] https://scop.berkeley.edu/downloads/?ver=2.08

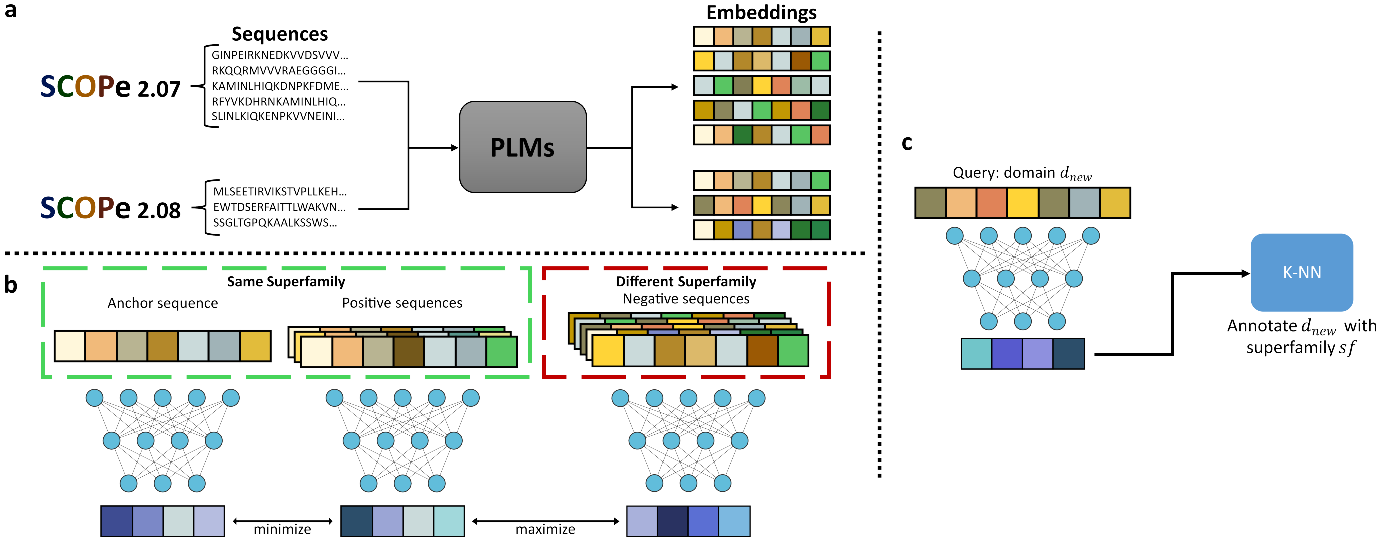


Fig. 1: Overview of the DomClass workflow. **(a)** Protein domain sequences from SCOPe 2.07 and 2.08 are embedded using PLMs. **(b)** A feedforward neural network is trained with a contrastive learning objective to refine the embedding space by pulling together domains from the same superfamily (positives) and pushing apart domains from different superfamilies (negatives). The resulting representation space – depicted by output vectors in shades of blue – is optimized to reflect evolutionary relationships. **(c)** Superfamily annotation is performed using a distance-weighted K-NN classifier in the embedding space: a query domain d_{new} is assigned to a superfamily sf based on its proximity to annotated domains in the embedding space.

thus providing a more challenging contrast [35]. The full contrastive set is defined as $A(d) = N(d) \cup P(d)$, and τ denotes a temperature scaling factor (set to 0.1).

The learning objective is a contrastive loss function that minimizes the distance between the anchor and the positives while maximizing the distance between the anchor and the negatives (Fig. 1b).

C. Annotation of Superfamilies

Once the embedding space has been trained to reflect evolutionary relationships, we annotate domains with superfamilies using a K-Nearest-Neighbors (K-NN) strategy to make predictions (Fig. 1c). This approach takes advantage of the geometric structure of the learned space to infer the class label of unseen domains without requiring additional training. Instead of treating all neighbors equally, we adopt a distance-weighted voting scheme [14], in which closer – and thus more similar – domains have a proportionally greater influence on the final prediction

This distance-weighted strategy is essential for handling the severe class imbalance in SCOPe: in version 2.07, roughly 40% of superfamilies have fewer than five representative domains, and 16% are represented by only a single domain. A uniform voting strategy would systematically bias predictions toward large, well-populated superfamilies.

For example, consider the case of $k = 3$, where two of the nearest neighbors belong to a majority superfamily and one to a minority superfamily represented by a single domain. Under uniform voting, the majority label would be selected, even if the two majority neighbors are farther from the query than the closer minority neighbor. In contrast, distance-weighted voting gives greater influence to the nearest neighbors, leading to more refined decision boundaries in the embedding space.

Given a query embedding z_q, the algorithm identifies the K closest training embeddings using Euclidean distance. The predicted superfamily label $\hat{y}$ is then given by:

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} \sum_{i=1}^{k} w_i \cdot \mathbb{I}(y_i = y), \quad w_i = \frac{1}{d(z_q, z_i)} \tag{2}$$

Here, $\mathcal{Y}$ denotes the set of all possible superfamily labels, y_i represents the superfamily label of the i-th nearest neighbor, and $\mathbb{I}(y_i = y)$ is an indicator function that equals 1 if the i-th neighbor belongs to class y, and 0 otherwise. The weight w_i of each of the k nearest neighbors z_i is inversely proportional to their Euclidean distance from the query embedding.

IV. Experimental Setup

We frame the problem of identifying remote homology as a multi-class classification problem where the goal is to predict SCOPe superfamilies. To construct a challenging validation set, we sample 10% of domains from each superfamily in the training set, ensuring that no selected domain shares more than 40% sequence identity with any other domain in the dataset. This approach promotes sequence diversity and is commonly used to assess remote homology detection [9], [11]. Therefore, we train our models with 90% of the domains in SCOPe 2.07, and the remaining 10% are used to select model parameters.

We conduct our experiments in a prospective setting, using the latest version of SCOPe 2.08. This strategy simulates a real-world scenario where we can test whether a classification method could have been effective in assigning superfamilies to new domains when only SCOPe 2.07 was available. By preserving the temporal structure of the data, the evaluation mirrors the conditions under which domain annotation would be made in practice, thereby providing a realistic assessment of the models' utility [8], [18]. We constructed this prospective test set by retrieving the new domains in SCOPe 2.08 that belong to superfamilies already present in SCOPe 2.07.

To assess the applicability of protein embeddings to this task, we implemented and tested the following methods:

- **DOMCLASS-1** is our contrastive learning approach with K-NN classification, defined in the *Methods* section. The input layer corresponds to the ESM-1b protein embeddings with a dimension of $1,280$. This is followed by two hidden layers of $2,048$ neurons each and ReLU activation, and a final output layer of 256 neurons. Superfamilies are predicted using K-NN as defined in Eq. 2, with $k = 3$.
- **DOMCLASS-2** is a variant of our approach that uses the ESM2-3B protein embeddings. The input layer corresponds to the protein embeddings with a dimension of $2,560$., followed by three hidden layers of $1,024$ neurons and ReLU activation, and a final output layer of 256 neurons. Similarly, superfamilies are predicted using K-NN as defined in Eq. 2, with $k = 3$.
- **ESM-1b-NN** is a feedforward neural network trained to directly predict SCOPe superfamilies as a multi-class classification task. It takes the frozen protein embeddings as input and outputs a probability distribution over all superfamilies. This model serves as a straightforward baseline to evaluate how well the original embedding space supports classification, without additional supervision on the embedding structure. It also provides a point of comparison for DOMCLASS, allowing us to assess whether contrastive learning leads to better-aligned embeddings for remote homology detection. This model consists of an input layer of $1,280$ neurons corresponding ot the ESM-1b protein embeddings. This is followed by two hidden layers of $1,500$ neurons each and ReLU activation, and a final output layer of $2,006$ neurons and softmax activation.
- **ESM2-3B-NN** is a variation of ESM-1b-NN where the main difference is that we use the ESM2-3B embeddings as the input to the feedforward neural network. It consists of an input layer of $2,560$ neurons, followed by two hidden layers of $1,500$ neurons each and ReLU activation, and a final output layer of $2,006$ neurons and softmax activation.
- **ESM-1b-raw** serves as a baseline to evaluate the predictive signal present in the unmodified protein embeddings generated by the general-purpose ESM-1b model. Without any task-specific fine-tuning or supervised training, we apply the K-NN strategy (Eq. 2), with $k = 3$, directly on these embeddings to assign superfamilies. Prior work has shown that ESM-1b embeddings reflect a degree of structural relatedness which can be informative of protein homology [29]. This baseline allows us to assess whether the pre-trained embedding space inherently supports remote homology detection.
- **ESM2-3B-raw** is equivalent to **ESM-1b-raw** but uses embeddings generated by the larger ESM2-3B model. We apply the same K-NN strategy (Eq. 2), with $k = 3$, directly on these embeddings to assign superfamilies.
- **BLASTp** serves as a conventional baseline to evaluate how well standard sequence-based similarity approaches perform in remote homology detection [24], [35]. In our setup, we use the test set as the query and the training set as the subject database. For each query domain, we assign the superfamily label of the top hit based on the highest percent sequence identity.
- **HMMER** provides a profile-based baseline to assess how well probabilistic models of sequence superfamilies capture remote homology [15], [19]. In our setup, we construct one HMM profile per superfamily in SCOPe 2.07. We then use the test set as the input sequences and search them against the HMM database. Each test domain is assigned the superfamily label of the top-scoring profile based on the highest bit score.

To evaluate the effectiveness of the different methods in the task of predicting SCOPe superfamily annotations, we employ a set of robust performance metrics commonly used in the multi-class scenario [3], [19], [32]:

1) **Recall (weighted)**: This measures the proportion of correctly predicted instances among all instances of a given class.

$$\begin{aligned} \text{Recall} &= \sum_{k=1}^{K} w_k \cdot \frac{TP_k}{TP_k + FN_k} \\ w_k &= \frac{TP_k + FN_k}{\sum_{j=1}^{K}(TP_j + FN_j)} \end{aligned} \tag{3}$$

where TP_k and FN_k denote the number of true positives and false negatives for each class k, and the weight w_k ensures that more frequent classes contribute proportionally to the overall score.

2) **Matthews Correlation Coefficient (MCC)**: MCC is a correlation coefficient between predicted and true labels. It generalizes to multi-class classification by operating over the entire confusion matrix. [1].

$$\text{MCC} = \frac{c \cdot s - \sum_k p_k t_k}{\sqrt{(s^2 - \sum_k p_k^2)(s^2 - \sum_k t_k^2)}} \tag{4}$$

where $c = \sum_k C_{kk}$ is the total number of correctly classified instances (diagonal of the confusion matrix), $p_k = \sum_j C_{kj}$ is the total number of predicted instances for class k, $t_k = \sum_i C_{ik}$ is the total number of true instances of class k, and $s = \sum_{i,j} C_{ij}$ is the total number of instances.

All metrics were calculated treating the multi-class structure to provide an overall view of model performance. This allows a fair comparison between models in the presence of highly unbalanced class distributions typical of superfamily classification.

V. Results

The results are organized into three main comparisons. First, we evaluate whether ESM-1b-raw and ESM2-3B-raw can outperform BLASTp and HMMER. Next, we assess whether using ESM-1b-NN and ESM2-3B-NN improves predictive accuracy over their raw alternatives. Finally, we examine whether our contrastive learning methods DOMCLASS-1 and DOMCLASS-2 lead to further gains over both the raw and classification-based approaches.

To evaluate performance under varying levels of sequence similarity, we report results on three test sets. The first, referred to as "Full", contains all domains in our test set without filtering based on sequence identity (16.222 domains from 782 superfamilies). The other two, "40%" and "30%", are subsets of "Full" that include only domains with less than 40% and 30% sequence identity to any domain in the training or test sets. These subsets contain 930 domains from 175 superfamilies and 61 domains from 22 superfamilies, respectively. The "30%" set corresponds to the *twilight zone* of sequence similarity [30], where detecting homology becomes particularly challenging. These subsets allow us to assess model performance under increasingly difficult remote homology scenarios.

Our experimental results are summarized in Figures 2 (a), and (b), which report recall and MCC, respectively, for all methods across the different test sets.

A. Protein embeddings can go beyond sequence

The results of our prospective evaluation indicate that the PLM embeddings generated by ESM-1b and ESM2-3B are informative of evolutionary relationships. The relative distances between domain embeddings already allow for effective remote homology detection using a simple K-NN classifier, without any supervised refinement (ESM-1b-raw and ESM2-3B-raw in Fig. 2). The superior performance compared to BLASTp and HMMER, especially in the low sequence identity subsets, suggests that the embedding distances in both ESM spaces capture evolutionary features that go beyond what is represented by sequence similarity alone.

Notably, ESM2-3b-raw performs consistently below ESM-1b-raw across all test sets, with the performance gap increasing in the 30% identity subset. This is consistent with the findings of Rives et al. [29], who report that ESM-1b embeddings organize sequences according to remote homology and reflect structural features. The high performance observed in the "Full" set is expected, as domains with similar sequences tend to be embedded close together, given that the ESM models were pre-trained on hundreds of millions of protein sequences.

B. Supervised classification using protein embeddings

We evaluated whether incorporating protein embeddings into a supervised classification framework could improve prediction accuracy over using raw embeddings directly. This expectation holds for ESM2-3B-NN, which consistently outperforms ESM2-3B-raw across all test sets.

However, this improvement is not consistent across models. Both ESM2-3B-NN and ESM-1b-NN are generally less accurate than ESM-1b-raw. This can be explained, in part, by the fact that distances between ESM-1b domain embeddings are more informative about evolutionary relationships than those from ESM2-3B. As a result, it is more difficult for a supervised model to outperform ESM-1b-raw than to outperform ESM2-3B-raw.

C. Contrastive learning improves remote homology detection

Our proposed DOMCLASS models outperform all other approaches, including ESM-1b-raw. This indicates that ESM-1b domain embeddings also encode evolutionary features that are not fully reflected in the raw embedding distances. Unlike ESM-1b-NN, the contrastive learning strategy employed by DOMCLASS-1 can effectively exploit these features by obtaining a more refined embedding space specially suited for domain superfamily prediction.

Interestingly, DOMCLASS-2 achieves the best overall performance, despite ESM2-3B-raw performing worse than ESM-1b-raw. This suggests that the poorer performance of ESM2-3B-raw is not due to a lack of relevant information, but to the distances in the embedding space being less well-aligned for direct similarity-based classification. The contrastive learning framework is able to refine this space to reflect evolutionary relationships more clearly. Thus, our contrastive learning framework proves effective in transforming pre-trained PLM embeddings into a representation space where evolutionary relationships are more explicitly reflected in the distances.

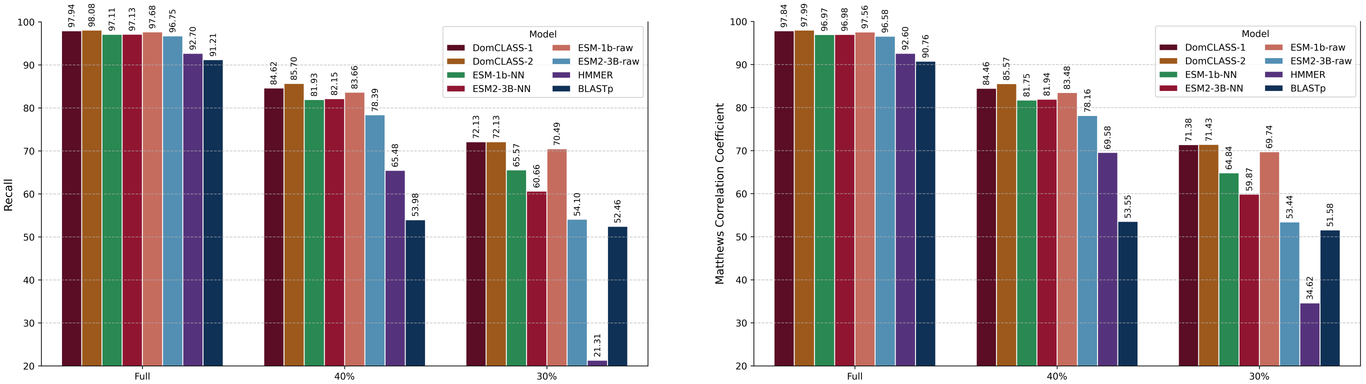
VI. DISCUSSION

SCOPe relies on expert manual curation to group domains with strong structural and functional evidence of common evolutionary origin [25]. Our aim is to analyze whether domain embeddings, obtained from PLMs pretrained with millions of sequences, can be used to make predictions about these SCOPe superfamilies— i.e, whether they may encode the features that experts would use to assess homology.

Our analysis reveals that the distances between domain embeddings from ESM-1b and ESM2-3B can uncover remote homologs that standard sequence identity measures and HMM-based searches fail to detect, particularly in the twilight zone of sequence identity [30]. Furthermore, we demonstrate that improving prediction performance using a simple neural network trained on these pre-trained embeddings (ESM-NN) is not straightforward, highlighting the challenges of leveraging the embeddings directly.

Furthermore, we proposed DOMCLASS, a contrastive learning approach that can effectively exploit evolutionary features present in the pre-trained ESM embeddings that are not fully reflected in the embedding distances alone. DOMCLASS models are trained in a supervised fashion to project domain embeddings into a lower-dimensional space in which domains from the same superfamily cluster together – improving homology prediction, as confirmed by our experiments.

We also note that using domain embeddings from different PLMs, such as other models from the ESM family, to train new DOMCLASS models can incorporate more nuanced features that can improve predictive prowess. In particular, the main ESM2 model, which has 15 billion parameters and has been



(a) Recall

(b) Matthews Correlation Coefficient

Fig. 2: Performance comparison of different evaluation metrics across different sequence similarity thresholds: Recall (left) and MCC (right).

shown to outperform ESM2-3B on downstream tasks [23], could be a good candidate.

Besides training with more sophisticated PLMs embeddings, for future work, we plan to improve the selection of negative examples for our contrastive learning framework using the SCOPe hierarchy itself. While SCOPe superfamilies reflect evolutionary relationships, broader groupings such as Folds represent domains with similar structural features but not necessarily shared ancestry. We believe that selecting negatives from the same Fold, but different superfamilies, could lead to a more fine-grained and evolutionarily meaningful embedding space.

VII. Acknowledgments

References

[1] Comparing two k-category assignments by a k-category correlation coefficient. *Computational biology and chemistry*, 28(5-6):367–374, 2004.

[2] Bruce Alberts, Rebecca Heald, Alexander Johnson, David Morgan, Martin Raff, Keith Roberts, and Peter Walter. *Molecular biology of the cell: seventh international student edition with registration card*. WW Norton & Company, 2022.

[3] José Juan Almagro Armenteros, Casper Kaae Sønderby, Søren Kaae Sønderby, Henrik Nielsen, and Ole Winther. Deeploc: prediction of protein subcellular localization using deep learning. *Bioinformatics*, 33(21):3387–3395, 2017.

[4] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[5] Tristan Bepler and Bonnie Berger. Learning protein sequence embeddings using information from structure. In *International Conference on Learning Representations*, 2019.

[6] Carl Ivar Branden and John Tooze. *Introduction to protein structure*. Garland Science, 2012.

[7] Nadav Brandes, Grant Goldman, Charlotte H. Wang, Chun Jimmie Ye, and Vasilis Ntranos. Genome-wide prediction of disease variant effects with a deep protein language model. *Nature Genetics*, 55(9):1512–1522, August 2023.

[8] Aurel Cami, Alana Arnold, Shannon Manzi, and Ben Reis. Predicting adverse drug events using pharmacological network models. *Science translational medicine*, 3(114):114ra127–114ra127, 2011.

[9] John-Marc Chandonia, Gary Hon, Nigel S Walker, Loredana Lo Conte, Patrice Koehl, Michael Levitt, and Steven E Brenner. The astral compendium in 2004. *Nucleic acids research*, 32(suppl_1):D189–D192, 2004.

[10] Junjie Chen, Mingyue Guo, Xiaolong Wang, and Bin Liu. A comprehensive review and comparison of different computational methods for protein remote homology detection. *Briefings in bioinformatics*, 19(2):231–244, 2018.

[11] Junjie Chen, Ren Long, Xiao-long Wang, Bin Liu, and Kuo-Chen Chou. drhp-psera: detecting remote homology proteins using profile-based pseudo protein sequence and rank aggregation. *Scientific reports*, 6(1):32333, 2016.

[12] Davide Chicco, Valery Starovoitov, and Giuseppe Jurman. The benefits of the matthews correlation coefficient (mcc) over the diagnostic odds ratio (dor) in binary classification assessment. *Ieee Access*, 9:47112–47124, 2021.

[13] Andrew Dickson and Mohammad RK Mofrad. Fine-tuning protein embeddings for functional similarity evaluation. *Bioinformatics*, 40(8):btae445, 2024.

[14] Sahibsingh A Dudani. The distance-weighted k-nearest-neighbor rule. *IEEE Transactions on Systems, Man, and Cybernetics*, (4):325–327, 1976.

[15] Sean R Eddy. Accelerated profile hmm searches. *PLoS computational biology*, 7(10):e1002195, 2011.

[16] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rihawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. Prottrans: Towards cracking the language of life's code through self-supervised deep learning and high performance computing, 2021.

[17] Naomi K Fox, Steven E Brenner, and John-Marc Chandonia. Scope: Structural classification of proteins—extended, integrating scop and astral data and classification of new structures. *Nucleic acids research*, 42(D1):D304–D309, 2014.

[18] Diego Galeano, Shantao Li, Mark Gerstein, and Alberto Paccanaro. Predicting the frequencies of drug side effects. *Nature communications*, 11(1):4575, 2020.

[19] Michael Heinzinger, Maria Littmann, Ian Sillitoe, Nicola Bordin, Christine Orengo, and Burkhard Rost. Contrastive learning on protein embeddings enlightens midnight zone. *NAR genomics and bioinformatics*, 4(2):lqac043, 2022.

[20] Richard Hughey and Anders Krogh. *SAM: Sequence alignment and modeling software system*. University of California at Santa Cruz, 1995.

[21] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

[22] Anders Krogh, Michael Brown, I Saira Mian, Kimmen Sjölander, and David Haussler. Hidden markov models in computational biology: Applications to protein modeling. *Journal of molecular biology*, 235(5):1501–1531, 1994.

[23] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, March 2023.

[24] Wei Liu, Ziye Wang, Ronghui You, Chenghan Xie, Hong Wei, Yi Xiong, Jianyi Yang, and Shanfeng Zhu. Plmsearch: Protein language model powers accurate and fast sequence search for remote homology. *Nature communications*, 15(1):2775, 2024.

[25] Loredana Lo Conte, Bart Ailey, Tim JP Hubbard, Steven E Brenner, Alexey G Murzin, and Cyrus Chothia. Scop: a structural classification of proteins database. *Nucleic acids research*, 28(1):257–259, 2000.

[26] Kevin Molloy, M Jennifer Van, Daniel Barbara, and Amarda Shehu. Exploring representations of protein structure for automated remote homology detection and mapping of protein structure space. *BMC bioinformatics*, 15:1–14, 2014.

[27] Muhammad Khaerul Naim, Tati Rajab Mengko, Rukman Hertadi, Ayu Purwarianti, and Meredita Susanty. Embedcaps-dbp: Predicting dna-binding proteins using protein sequence embedding and capsule network. *IEEE Access*, 11:121256–121268, 2023.

[28] Christine A Orengo, Alex D Michie, Susan Jones, David T Jones, Mark B Swindells, and Janet M Thornton. Cath–a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1109, 1997.

[29] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

[30] Burkhard Rost. Twilight zone of protein sequence alignments. *Protein engineering*, 12(2):85–94, 1999.

[31] Johannes Söding. Protein homology detection by hmm–hmm comparison. *Bioinformatics*, 21(7):951–960, 2005.

[32] Felix Teufel, José Juan Almagro Armenteros, Alexander Rosenberg Johansen, Magnús Halldór Gíslason, Silas Irby Pihl, Konstantinos D Tsirigos, Ole Winther, Søren Brunak, Gunnar von Heijne, and Henrik Nielsen. Signalp 6.0 predicts all five types of signal peptides using protein language models. *Nature biotechnology*, 40(7):1023–1025, 2022.

[33] Michel Van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron LM Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with foldseek. *Nature biotechnology*, 42(2):243–246, 2024.

[34] Lan Xu. Deep learning for protein-protein contact prediction using evolutionary scale modeling (esm) feature. In *International Artificial Intelligence Conference*, pages 98–111. Springer, 2023.

[35] Tianhao Yu, Haiyang Cui, Jianan Canal Li, Yunan Luo, Guangde Jiang, and Huimin Zhao. Enzyme function prediction using contrastive learning. *Science*, 379(6639):1358–1363, 2023.